Tatjana Marvin[*]
University of Ljubljana
Jure Derganc[**]
University of Ljubljana
Saba Battelino[***]
University Medical Centre Ljubljana, University of Ljubljana

UDK 811.163.6'373:612.85
DOI: 10.4312/linguistica.57.1.197-210

ADAPTING THE FREIBURG MONOSYLLABIC WORD TEST FOR SLOVENIAN

1 INTRODUCTION

Speech audiometry is one of the standard methods used to diagnose the type of hearing loss and to assess the communication function of the patient by determining the level of the patient's ability to understand and repeat words presented to him or her in a hearing test. For this purpose, the Slovenian adaptation of the German tests developed by Hahlbrock (1953, 1960) – the Freiburg Monosyllabic Word Test and the Freiburg Number Test – are used in Slovenia (adapted in 1968 by Pompe). These tests employ the use of phonetically balanced lists of existent monosyllabic words with the aim of determining the percentage of correctly repeated words at different sound intensity levels.

In this study we focus on the Freiburg Monosyllabic Word Test for Slovenian, which has been criticized by patients in personal communication during and after testing, as well in the literature for the unequal difficulty and frequency of the words, with many of extremely rare or even obsolete items (Podlesek et al. 2007; Podlesek et al. 2008).[1] As part of the patient's communication function is retrieving the meaning of individual words by guessing, the frequency of use of an individual word is crucial. The less frequent and consequently less familiar words (e.g. *dac*, *golč*, *irh*, *lat*, *raš*, *sak*) do not contribute to a reliable testing result, as they cannot be guessed to the same extent as more familiar words (e.g. *bor*, *klop*, *pas*, *sin*). We therefore propose that the test be adapted by identifying and removing less familiar words from the list and supplementing them with phonetically similar words so as to preserve the phonetic balance of the list.

The paper is organized as follows. In Section 2 we provide a general description of the Freiburg test. In Section 3 we proceed to identify less familiar words in the Freiburg

[*] tatjana.marvin@ff.uni-lj.si

[**] jure.derganc@mf.uni-lj.si

[***] saba.battelino@mf.uni-lj.si

[1] In Podlesek et al. (2007) the authors criticize the Freiburg test from 1968 on the same grounds as in this paper, but with a different purpose. The authors eliminate a list of less familiar words, keeping only 135 items, and develop a different method of testing (the so-called staircase method), which, however, did not succeed in everyday clinical use due to a difficulty in comparing its results to those of the Freiburg test and the data from the relevant literature.

test, while in Section 4 we describe the procedure of replacing the less familiar words with more familiar ones that we extract from various Slovenian corpora, the result being a new version of the Freiburg test. Section 5 discusses some remaining issues and Section 6 concludes the paper.

2 FREIBURG MONOSYLLABIC WORD TEST FOR SLOVENIAN

2.1 Speech Audiometry

Classic pure-tone audiometry assesses only basic deficits in auditory function. The audiometric curve determines the detected threshold levels (in dB) for selected frequencies. To evaluate the clinical impact of hearing loss, disorders affecting auditory pathways after the cochlea, and especially the rehabilitation of severe hearing loss and deafness with cochlear implants, it is necessary to use various other audiometric tests, such as sound localization, auditory discrimination, auditory pattern processing and speech audiometry. Speech audiometry assesses the understanding of words presented at a specified loudness in different conditions (Musiek et al. 2011).

A speech recognition test consists of the patient's listening and repeating words, with the clinician marking a tally of right and wrong responses. The percentage of test words correctly repeated by the patient is referred to as the speech recognition score (also the word recognition score or speech discrimination score). The percentage of the correctly repeated words depends on more than just the patient's speech recognition ability; it also depends on the patient's familiarity with the words and on the intensity at which the words are presented. The graph of performance – intensity function shows how the patient's speech recognition performance depends on the intensity of the test materials (Gelfand 2009).[2] Different diseases of the middle, inner ear and central auditory pathways result in different speech recognition scores and different performance – intensity curves, despite having similar pure tone audiometry curves (van Dijk et al. 2000). The speech recognition score is essential in evaluating and comparing the rehabilitation effects achieved with the use of classical hearing aids or cochlear implants (De Riuter 2015).

In the Department of Otorhinolaryngology, University Medical Centre Ljubljana, speech is normally assessed with the Slovenian adaptation of the German tests developed in Hahlbrock (1953, 1960), and the Freiburg Number Test. The Slovenian adaptations of these were developed in Pompe (1968). After the patient is fitted with a hearing aid, a 20% improvement of the speech recognition score represents a significant therapeutic effect. However, if the score of a patient fitted with a standard hearing aid is lower than 50% of the highest possible score, then this approach is not a satisfactory rehabilitation method, and a cochlear implant should be considered. Moreover, a constantly improving score following the rehabilitation of a cochlear implant user is proof of a well-selected speech rehabilitation method and the effective work of the related language specialists.

[2] There exists speech audiometry in which sentences (rather than words) are used, but the correct repetition is in such tests much more influenced by other factors than in speech audiometry with monosyllabic words. Consequently, the results among the tested groups are not comparable.

2.2 Freiburg Monosyllabic Word Test

In this paper we focus on the Slovenian adaptation of the Freiburg Monosyllabic Word Test (henceforth Freiburg Test-SLO-1968). This consists of 281 monosyllabic nouns in the nominative singular form (nine of them repeated). In the test, a patient listens to phonetically balanced columns of 28–29 monosyllabic Slovenian words in a quiet environment, with the stimulus intensity level increased in each consecutive column. A speech audiogram with the percentage of correctly repeated words at each level serves as the basis for estimating the patient's communication function.

The Freiburg Test-SLO-1968 is phonetically balanced in the sense that the columns consisting of 28 or 29 words contain equal numbers of different letters (with rare exceptions), as can be seen in Table 1. For example, in each column there are nine occurrences of the letter "a," seven of the letter "o," four of the letter "g" and so on.[3]

Table 1: Letter frequency in the Freiburg Test-SLO-1968 in test columns 1 through 10.

	a	**b**	**c**	**č**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**r**	**s**	**š**	**t**	**u**	**v**	**z**	**ž**
1	9	2	2	3	5	4	1	4	3	5	1	4	5	3	4	7	6	10	8	1	8	2	4	1	1
2	9	2	1	3	5	4	0	4	3	4	1	4	5	3	4	7	5	10	9	1	8	2	4	1	1
3	9	2	1	3	5	4	0	4	3	4	1	4	6	3	4	7	5	10	9	1	8	2	3	1	1
4	9	2	1	3	6	5	0	4	3	4	1	4	5	3	4	7	6	11	9	1	7	2	4	1	1
5	9	2	1	3	5	4	1	4	3	4	1	4	5	3	4	7	4	10	9	1	8	2	4	1	1
6	9	2	1	3	5	4	0	4	3	4	1	4	6	3	4	7	5	10	9	1	8	2	3	1	1
7	9	2	1	3	5	5	1	4	3	4	1	4	6	3	4	7	5	10	10	1	9	2	3	1	1
8	9	3	1	3	5	4	0	4	3	4	1	4	6	3	4	7	4	11	9	1	8	2	3	1	1
9	9	2	1	3	5	4	0	4	3	4	1	4	4	3	4	7	5	10	9	1	8	2	5	1	1
10	9	2	1	3	6	5	0	4	3	4	1	4	4	3	5	7	5	11	9	1	8	2	5	1	1

The comparison of the occurrence of individual letters in the test to the occurrence of individual letters in the Slovenian language, as established in Jakopin (1999), reveals that the distribution of letters broadly reflects that in the actual language (Table 2). For example, the letter "f," which is rarely found in the language, appears in only three columns, while "m" and "t" appear in all of them, as their frequency of occurrence is much higher. Moreover, the letter "a" occurs nine times in each column, which is consistent with it being one of the most frequently occurring letters in the language. However, some discrepancies can be observed, as for example, the letter "j" occurs almost five times less often in the test than in the language, while the letter "f" occurs three times more often in the test (despite it being present in only three columns).

[3] In this paper we establish phonetic balance by referring to the letters in writing and not the actual sounds. See 5.1 for a discussion on the sound-letter relation.

Table 2: Letter frequencies in Slovenian literature (Jakopin 1999), in the Freiburg-SLO-1968, and the ratio between the two.

letter	Jakopin 1999 [%]	Freiburg-SLO-1968 [%]	ratio
a	10.5	8.9	0.85
b	1.9	2.1	1.07
c	0.7	1.1	1.64
č	1.5	3.0	2.00
d	3.4	5.1	1.51
e	10.7	4.2	0.40
f	0.1	0.3	2.69
g	1.6	3.9	2.40
h	1.1	3.0	2.81
i	9.0	4.0	0.45
j	4.7	1.0	0.21
k	3.7	3.9	1.07
l	5.3	5.1	0.97
m	3.3	3.0	0.90
n	6.3	4.0	0.64
o	9.1	6.9	0.76
p	3.4	4.9	1.46
r	5.0	10.1	2.03
s	5.1	8.9	1.76
š	1.0	1.0	0.99
t	4.3	7.9	1.82
u	1.9	2.0	1.05
v	3.8	3.7	1.00
z	2.1	1.0	0.47
ž	0.7	1.0	1.52

3 IDENTIFYING LESS FAMILIAR WORDS IN THE FREIBURG TEST

The first step in our project is to identify and eliminate the less familiar words that appear in the Freiburg Test-SLO-1968. As there is no existing data based on the patient judgements of the words' frequencies, we refer to Podlesek et al. (2007), the corpus of written Slovenian Gigafida, and the corpus of spoken Slovenian GOS.

3.1 Podlesek et al. (2007)

Podlesek et al. (2007) gathered data on the frequency of the Freiburg test words in everyday spoken language, as judged by the native speakers of Slovenian. The frequency was assessed by a sample of 141 students who were given written lists of the Freiburg test words and asked to assess the frequency of occurrence of each in their everyday lives (i.e. how often they hear it on TV, radio, or use it in spoken language) by using a 5-point Likert scale (0 – never, 5 – very often). The information gathered is one of our criteria in providing a list of less familiar words. For a word to be considered less familiar, we set the threshold at the average score 1 or less (on a scale from 1 to 5) in the Podlesek et al. (2007) survey of native speakers. There are 65 such words, listed in (1).

(1) *ar*, *ceh*, *cep*, *cis*, *čad*, *črm*, *dac*, *dis*, *dož*, *drač*, *dreg*, *dvir*, *gat*, *gnjat*, *golč*, *golk*, *golt*, *gož*, *grod*, *groh*, *hrst*, *il*, *irh*, *jad*, *jam*, *karp*, *krc*, *krm*, *krn*, *lat*, *loč*, *lug*, *mig*, *mik*, *nrav*, *or*, *pah*, *pard*, *plač*, *polk*, *ral*, *raš*, *rig*, *ril*, *rovt*, *sak*, *sekt*, *ser*, *sip*, *skrak*, *sna*, *sned*, *snet*, *soj*, *speh*, *spuh*, *stog*, *stud*, *svest*, *sviž*, *šeh*, *tvar*, *urh*, *vat*, *žad*

A vast majority of the words in (1) are completely unknown to contemporary native speakers of Slovenian, many of them being archaic terms relating to agricultural practices, animals and plants.

3.2 Reference Corpora

Several reference corpora for the Slovenian language are freely accessible at the internet portal of the project "Communication in Slovene" (http://eng.slovenscina.eu/korpusi). A natural choice for determining the frequency of spoken words would be the corpus of spoken Slovenian GOS (for details on the corpus see Zemljarič Miklavčič et al. 2009; Verdonik et al. 2013). The corpus contains transcripts of approximately 120 hours of speech found in various situations: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations. All speech is transcribed in two versions – with pronunciation-based spelling and with standardized spelling. The corpus contains around one million words. However, the drawback of the corpus is that it is still relatively small and many of the words that we would expect in a corpus of Slovenian are not found there, or have a very low number of hits (*ceh* "guild"– 0, *noj* "ostrich"– 0, *volk* "wolf"- 4).

A much larger corpus than GOS is Gigafida (an upgrade of Fidaplus), which contains about 1.2 billion words (see Erjavec and Logar Berginc (2012), Logar Berginc and Krek (2012) and Logar Berginc et al. (2012) for more information on this corpus). The corpus has been automatically lemmatized and includes morphosyntactic descriptions (part-of-speech, gender, case, number). The option "advanced search" enables the user to determine the part-of-speech of the word (noun, verb, adjective) as well as choose whether to search only for a particular form or for all forms of a word. This search engine is to some extent successful in eliminating the erroneous hits when the part-of-speech is specified. For example, when searching for the word *teč* "run-supine", where the option noun is chosen, the search gives no hits, which is expected, as *teč* is a verb. Yet, with *smuč* – in an advanced search specified for noun – the hits are all the noun, verb and adjective occurrences, as the search engine provides hits with the meaning "pike perch" (nouns) as well as those with the meaning "ski" (verbs and adjectives, e.g. *smuč. skoki* "ski jumps"). Even narrowing down the search to only the form *smuč* does not help – we still get the adjectival hits related to the meaning "ski". This is probably due to the incorrect connection between the lemma *smuč* "pike perch", present in the *Slovar slovenskega knjižnega jezika* (*Dictionary of Standard Slovenian*, hereafter *SSKJ*), and all the forms *smuč* – the ones relating to the meaning "pike perch" and those relating to the meaning "ski" (since the two are homonymous).

Reference corpora should thus be used with caution when establishing the frequency of individual words. Regardless of which corpus is used, special attention should be paid to the content of the results. There are numerous cases where an unlimited search in a corpus provides a very high number of hits for a certain word, but it then turns out that the vast majority of these are not for the word checked, but for some other, more familiar word that is homonymous with the word in question.

3.3 Final List of the Less Familiar Words in the Freiburg Test-SLO-1968

We now return to establishing the final list of the words to be eliminated from the Freiburg Test-SLO-1968, combining the results in Podlesek et al. (2007) and data from corpora available to us. First, there are some additional words that need to be considered for elimination from the Freiburg Test-SLO-1968, despite the fact that their frequency was not judged as below 1 in Podlesek et al.'s (2007) native speakers' test:

(2) *dna*, *hot*, *lišp*, *smuč*, *teč*

Let us begin with the words *hot* and *smuč*, which score 1.06 and 1.07, respectively, in the test for native speakers. According to *SSKJ*, the two words have the meanings *hot "*an interjection for a horse" and *smuč "*pike perch", which are words that are rarely used by speakers in their everyday lives. Both words were also erroneously recognized as frequent in the corpora, as a careful examination revealed that all or a great number of their hits are not for the actual dictionary meanings of the two words, but rather refer to the borrowed combination *hot dog* (for *hot*) and to the adjective *smučarski*, abbreviated as *smuč* "ski" (as described in the previous section)[4].

A different problem occurs with the words *dna* and *teč*, which score 1.06 and 1.13, respectively. One of these is not a nominative singular noun, as is true of other nouns in the Freiburg Test-SLO-1968 (*dna*), while the other is not a noun at all (*teč*).[5] An examination of the hits in the corpora reveals that the ones for *dna* are actually the plural nominative or accusative forms of the word *dno* "bottom" or the acronym DNA, and the hits for *teč* are the supine forms of the verb *teči* "to run". Finally, we decide to replace the word *lišp,* as it is marked archaic in *SSKJ* as well as in *Slovenski pravopis* (*Slovenian Orthography 2001*).

As to the list in (1), we decide to keep the eight words in (3) in the test, basing our decision on our native speakers' intuition as well as the frequency of these words in the Gigafida corpus, where we consider only genuine hits.[6]

[4] The word *hot* has no hits in GOS, while *smuč* has 30 hits, but none for its original meaning.

[5] In GOS the two have 1 and 0 hits. They are not found in the forms *dna* and *teč* in SSKJ. The reason why the two words appear in the test at all is thus unclear.

[6] We employ various strategies to ensure that the hits are indeed the words we are searching for and not homonymous words with different meaning. For some words, we use an advanced search,

(3) *ar*, *ceh*, *gnjat*, *gož*, *polk*, *soj*, *urh*, *vat*

Table 3: Corpus data for *ar*, *ceh*, *gnjat*, *gož*, *polk*, *soj*, *urh*, and *vat*

	Gigafida (accessed April 4, 2016)
ar "are"	1494 (for *arov*)
gnjat "ham"	965
gož "grass snake"	347
polk "regiment"	3711
urh "toad"	101
vat "watt"	2229
ceh "guild"	3648
soj "shine"	6274

The final list of 62 words that we decide to remove from the Freiburg test list is shown below:

(4) *cep*, *cis*, *čad*, *črm*, *dac*, *dis*, *dna*, *dož*, *drač*, *dreg*, *dvir*, *gat*, *golč*, *golk*, *golt*, *grod*, *groh*, *hot*, *hrst*, *il*, *irh*, *jad*, *jam*, *karp*, *krc*, *krm*, *krn*, *lat*, *lišp*, *loč*, *lug*, *mig*, *mik*, *nrav*, *or*, *pah*, *pard*, *plač*, *ral*, *raš*, *rig*, *ril*, *rovt*, *sak*, *sekt*, *ser*, *sip*, *skrak*, *smuč*, *sna*, *sned*, *snet*, *speh*, *spuh*, *stog*, *stud*, *svest*, *sviž*, *šeh*, *teč*, *tvar*, *žad*

4 CONSTRUCTING THE FREIBURG TEST-SLO-2016

The next step is supplementing the words in (4) with more familiar words so as to preserve the letter frequencies in each column of the original test (Table 1). We begin by building the database of possible replacements (Section 4.1) and proceed to finding the optimal ones (Section 4.2).

4.1 Constructing the Database of Possible Replacement Words

To construct the database of possible replacements for the words in (4) we again refer to the corpora. For this purpose, we use the GOS and ccGigafida corpora, as they have XML sources available. The ccGigafida is ten times smaller than its base corpus Gigafida, and was made by random paragraph selection. The full Gigafida is unfortunately not available with its source, and ccGigafida is currently one of the largest freely available corpora of Slovenian (cc stands for the Creative Commons-Attribution-NonCommercial license).

where we specify the gender, thus eliminating the homonymous results (e.g. *polk* "regiment-masculine" is homonymous with *polk* "polka dance-feminine/plural/genitive"). For others, where this approach does not work, we search for a particular form of the word. For example, with the word *ar* "are" we search for *arov* "are-plural/genitive", as *ar* "are-singular/nominative" mostly gives completely unrelated hits (acronyms, *ara* "downpayment", etc.).

The database of monosyllabic singular nouns from GOS and ccGigafida was constructed by first extracting the lemmas of all nouns (by searching for the lemmas with the XML msd tag "S*", where * represents any number of any characters). We then extract all nouns with one vowel and all those that contain the sonorant *r*, but no vowel (as in these the schwa appears in pronunciation, but not in writing, e.g. *vrt* "garden"). The processing of the words in corpora was performed using Mathematica software (Wolfram Research), which has numerous built-in word-analysis tools. ccGigafida gives 21,942 hits and GOS 1,190. We then set the limit as to the number of hits for a word to be kept in the database; for the nouns from ccGigafida we set the limit at 200 hits, while we keep all the nouns from GOS, as this is a relatively small, but representative corpus of spoken language. We use both corpora because not all the words present in GOS have more than 200 hits in ccGigafida (e.g. *jež* "hedgehog" has eight hits in GOS and 187 in ccGigafida). This leaves us with 1,771 nouns from both corpora.

We then need to eliminate unsuitable words from this list. First, we exclude all the words that are acronyms or non-Slovenian words (*mr*, *oš*, *fahr*, *boys*, etc.). We then pass the words through another filter, as we need to eliminate colloquial or slang words (*šiht*, *starš*, *ksiht*, *kšeft*, *baš*, *dec*, etc.), the remaining English words (*show*, *what*, *fan*, *bird*, *pub*, etc.), acronyms (*dag*, *kfor*, *kud*, *pef*, *sos*, etc.), pronouns (*jaz*), proper names (*Ptuj*, *Cre*s, *Krim*, *Jan*, etc.), words potentially uncomfortable for the speaker to say (*seks*), vulgar words (*rit*, *fuk*, *drek*, etc.) and the words that have one vowel, but are in fact bisyllabic (*črka*, *brlog*, *črnec*, *prvak*, etc.). When this is completed, the remaining words have to be checked against the list in the Freiburg Test-SLO-1968, as we have to make sure not to use words already present in the test as replacements. The final list contains 348 monosyllabic nouns that are suitable as replacements.

4.2 Finding Optimal Replacements

The goal was to replace the less familiar words in each column of the Freiburg Test-SLO-1968 (4) with words from the list of suitable replacements that was derived from the corpora, while preserving the letter frequencies in each column. We rely on computational algorithms to find the combinations of words that satisfy the letter frequency criterion. This is a computationally demanding task, since for large word sets it is impossible to search through all possible combinations of words and test them for the desired criterion (for example, the number of all possible combinations of 10 words from a set of 348 words is on the order of 10^{20}). We therefore employ a version of a recursive back-tracking algorithm (Knuth 2016), where the paths that cannot lead to a solution are discarded from the search tree (e.g. if one is looking for two words with a total of five letters, there is no need to search through combinations of words with a larger number of letters).

To further speed up the computation, the words were first vectorized in a 25-dimensional space of letters according to their letter count, and the task was then solved in Matlab, which is optimized for efficient vector computation. As an illustration: if the alphabet contained only three letters, the unit vectors would be "a"= (1,0,0), "b"=(0,1,0), and "c"=(0,0,1), and the word "aaac" would, for example, correspond to a vector (3,0,1).

By using the optimized algorithm, we were able to find many combinations of the replacement words that exactly matched the letter count criterion for all columns except for column 5. For this column, where only two words had to be replaced (*ril* and *šeh*), the most suitable match differed by three letters.[7] Finally, the list of words that we propose for the Freiburg Test-SLO-2016 (shown in Table 4) was manually selected from the results returned by the algorithm.

Table 4: The proposed Freiburg Test-SLO-2016. The words that replaced less familiar words from the original test are marked with *. All the columns, except for column 5, exactly match the letter frequencies of the corresponding columns from the Freiburg Test-SLO-1968, while column 5 differs by three letters ("m," "v", "v").

1	2	3	4	5	6	7	8	9	10
lak	čir	jež	lan	bon	noj	tat	car	gož	niz
mah	paž	gad	kip	seč	sin	čer	rob	cev	del
vir	grm	piš	set	kad	mah	mož	kih	mir	laž
dlan	sla	beg	brc	srh	gon	ceh	sneg	voh	prag
prod	vid	svat	past	moč	dar	soj	vrat	trst	snov
tast	dom	park	zvon	dir	grah	film	žolč	disk	gumb
kljun	štor	molk	breg	cvet	most	kost	plen	vamp	tisk
sok	pest	stolp	dolg	bron	brat	svet	grom	trup	šport
grušč	noht	tun	smeh	gams	dvom	vrač	strah	vdih	vat
stvar	bran	hrast	trušč	vrag	plašč	strop	pad	zglob	blesk
lift	stran	stric	sad	pisk	kup	dan	zvrst	lord	strok
stih	log	post	vrisk	smrad	drozg	zdrob	blišč	pust	stan
gost	ključ	grič	slast	kal	strel	kramp	test	noč	dren
bog	vzrok	klop	punč	gnjat	glad	glas	vran	gnoj	urh
polh	rast	sod	vlak	tresk	vest	brst	drob	prah	dvor
bas	hrib	last	rep	vzor	slak	trg	mast	mag	kap
trn	svak	čoln	čast	slap	hči	nos	polk	krt	vrč
som	med	drn	nart	polž	polt	uk	nart	peč	os
zid	dvig	zob	maj	vrt	prst	svet	ud	bor	*ptič
reč	tank	dih	srd	list	čar	*sklad	*jed	sad	*cent
ep	pas	vas	pih	drog	ton	*čin	*kri	*član	*shod
*cmok	les	meh	vod	fant	ris	*dol	*laks	*rast	*čaj
*ranč	srp	ar	tla	čut	led	*ring	*las	*dres	*spis
*prt	*čas	*črv	rep	duh	bar	*dah	*hrt	*takt	*gram
*grad	*gol	*rang	*krom	pot	up	*greh	*čip	*sen	*rov
*dež	*hec	*dur	*hod	as	*cvek	*par	*gos	*sir	*gred
*prav	*top	*smer	*rog	*lev	*stik	*šal	*sum	*šal	*hlad
*vic	*trud	*tlak	*rod	*miš	*srž	*plus	*god	*kvas	*tram
*pes			*žig	*vrh		*tip			

[7] There was one replacement differing only by two letters, *pirh* "Easter egg" and *fleš* "flash." We decide against this option, as it contains the letter "f," which occurs in the Freiburg Test-SLO-1968 three times more often than in the language, and adding another example with this letter would further increase its frequency of occurrence.

5 SOME REMAINING ISSUES

5.1 Phonetic Balance

In this paper we follow Pompe and establish phonetic balance by referring to the letters in writing and not the actual sounds pronounced in the words. The phonetic balance achieved in this way is an approximation of the phonetic balance that takes into account the actual pronunciation. To explain this we need to refer to the notion of phoneme and allophone, and their relation to the letters in the alphabet. A phoneme is standardly defined as the smallest sound unit that can be segmented from the acoustic flow of speech and which functions as a semantically distinctive unit: if a sound unit is replaced by another sound unit in a word and the two words have a different meaning, we define the two sound units as phonemes, e.g. in the English pair pet – bet, /p/ and /b/ are phonemes. Phonemes are abstract units, each phoneme representing a class of phonetically similar sound variants, the allophones, which are in a complementary distribution, depending on the phonological environment they appear in. For example, in English, the phoneme /p/ has an aspirated variant [p^h] at the beginning of the syllable (as in *pet*), but a non-aspirated variant [p] elsewhere (e.g. *loop*).

The writing systems that use letters can be organized in different ways – some of them tend to use a letter to denote a phoneme, others are closer to using a letter for an allophone. In Slovenian, the tendency is for one letter to represent one phoneme. For example, the letter "n" stands for the phoneme /n/, which has three allophones: [ŋ] when followed by a velar consonant as in *Anglija* "England;"; [n^j] (for some speakers) when followed by [j#] or [jC] as in *konj* "horse," *konjski* "horse-adj" and [n] elsewhere, e.g. *nos* "nose". However, there is no one-to-one correspondence among phonemes and letters, as there are more phonemes than letters (29 versus 25). In fact, there are many cases in which a single letter stands for two or more phonemes, e.g. the letter "e" can denote [e] in *led* "ice", [ɛ] in *žep* "pocket" or [ə] in *pes* "dog". Finally, for some phonemes, no letter is used: in many words that contain the consonant [r] and the vowel [ə], the vowel is pronounced, but not expressed in writing: *vrt* "garden," *smrt* "death," etc.

Referring to letters instead of phonemes or allophones is thus an approximation on two levels. First, the letter-phoneme correspondence is not always one-to-one, and second, even if it were, the phonemes themselves can refer to different sounds in pronunciation, i.e. their allophones (see /n/ above). Referring to the allophones in the phonetic balance calculation would require a much more thorough linguistic analysis. It would, for example also require considering all the phonological rules that take place in Slovenian, such as the final devoicing of voiced obstruents (the word *bog* "god" is pronounced the same way as *bok* "hip"), the changes that occur at word boundaries, and the like. Moreover, the search engines of the corpora are organized according to written and not spoken language, with the exception of GOS, which is too small to be the only representative corpus of the language (see also Section 3.2). As corpora such as ccGigafida are crucial for establishing the word frequencies and building the database of words that we need for adapting the Freiburg test, we use the letters as approximations of the actual sounds, bearing in mind the limitations that come with this.

Another issue relating to phonetic balance that remains a challenge in our future research is balancing the occurrence of individual letters in the test with the occurrence of individual letters in the Slovenian language. We mentioned in Section 2 that the distribution of letters in the Freiburg Test-SLO-1968 only broadly reflects the distribution in the actual language (with the latter established in Jakopin (1999)). The same is true for the Freiburg Test-SLO-2016, as the new test has been designed in such a way that it preserves the phonetic balance of the older version. A more exact balance still remains to be achieved.

5.2 Syllable Structure

The Freiburg Test-SLO-1968 and the Freiburg Test-SLO-2016 are balanced with respect to the number of letters in individual columns consisting of 28 or 29 words. A possible balance to consider in future work is that with respect to the types of syllables that appear in the language, some of which are exemplified in (5) with the related notation shown in (6).

(5) *um* “mind” → Vs
gol “goal” → oVs
ples “dance” → osVo
ring “ring” → sVso
sklad “fund” → oosVo

(6) Notation
V for vowel; spelled a, e, i, o, u
s for sonorant consonant; spelled m, n, v, j, l, r
o for obstruent consonant spelled p, t, b, d, k, g, h, f, c, č, dž, s, š, z, ž

The analysis of syllable structure in the test columns shows that 36 different syllable combinations are used in the test. Given the fact that one column consists of 28 or 29 monosyllabic nouns, not all combinations can be present in each individual column. The balance in terms of syllable structure appears a complex issue, and we thus leave it for future research.

6 CONCLUSION

In this paper we adapted the 1968 version of the Freiburg Monosyllabic Word Test for Slovenian by identifying and removing less familiar words from the list, supplementing them with phonetically similar words so as to preserve the phonetic balance of the list. The result is a new test, the Freiburg Test-SLO-2016, as well as a new database of monosyllabic nouns that are commonly used by native speakers of Slovenian. The new Freiburg test presents a great improvement in speech audiometry clinical practice in Slovenia, while the related database can be used for constructing new tests for diagnosing hearing loss in the future. The new test with its clinical implementation will provide a tool for better assessment of the patient’s rehabilitation with different hearing aids. It will enable clinicians to select good candidates for cochlear implantation, and to distinguish different pathologies in central auditory pathways.

The adaptation crucially required the use of Slovenian corpora, the written corpus Gigafida and the spoken corpus GOS. These were of great help when determining the frequency of the words that appear in the test and in the extraction of new nouns needed as replacements, though we did encounter some problems with lemmatization and morphosyntactic tagging. In this light, we strongly encourage further funding and research of advanced algorithms of Slovenian corpora (e.g. accurate automatic lemmatization), as such work would greatly advance the application of the corpora in studies of the Slovenian language, resulting, among other things, in improved clinical practice with hearing impaired patients.

Acknowledgements
We would like to thank Simon Krek for his help with Slovenian corpora.

References

DE RUITER A. Mark/Joke A. DEBRUYNE/Michelene N. CHENAULT/Tom FRANCART/ Jan P. BROKX (2015) "Amplitude Modulation Detection and Speech Recognition in Late-Implanted Prelingually and Postlingually Deafened Cochlear Implant Users." *Ear Hear* 36/5: 557–566. http://dx.doi.org/10.1097/AUD.0000000000000162

ERJAVEC, Tomaž/Nataša LOGAR BERGINC (2012) "Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES." In: T. Erjavec/J. Žganec Gros (eds), *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Jožef Stefan Institute, 57–62.

HAHLBROCK, Karl Heinz (1953) "Über Sprachaudiometrie und neue Wörterteste." [On speech audiometry and new word tests]. *Arch Ohren Nasen Kehlkopfheilkd* 162: 394–431.

HAHLBROCK, Karl Heinz (1960) "Kritische Betrachtungen und vergleichende Untersuchungen der Schubertschen und Freiburger Sprachteste." [Critical reflection and comparative examination of the Schuberts and the Freiburg test]. *Zeitschrift für Laryngologie, Rhinologie, Otologie und Ihre Grenzgebiete* 39, 100.

JAKOPIN, Primož (1999) *Zgornja meja entropije pri besedilih v slovenskem jeziku*. Doctoral dissertation. University of Ljubljana.

KNUTH Donald E. (2016) *Introduction to backtracking. The Art of Computer Programming. Volume 4. Pre-fascicle 5B*. http://www.cs.utsa.edu/~wagner/knuth/

LOGAR BERGINC, Nataša/Miha GRČAR/Marko BRAKUS/Tomaž ERJAVEC/Špela ARHAR HOLDT/Simon KREK (2012) *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, FDV.

LOGAR BERGINC, Nataša/Simon KREK (2012) "New Slovene corpora within the communication in Slovene project." *Prace Filologiczne* 63, 197–207.

MUSIEK Frank E./Gail D. CHERMAK/Jeffrey WEIHING/Megan ZAPPULLA/Stephanie NAGLE (2011) "Diagnostic accuracy of established central auditory processing test batteries in patients with documented brain lesions." *Journal of the American Academy of Audiology* 22/6, 342–358. http://dx.doi.org/10.3766/jaaa.22.6.4

GELFAND, Stanley A. (2009) *Essentials of Audiology*. New York, Stutgart: Thieme.

PODLESEK, Anja/Luka KOMIDAR/Gregor SOČAN/Boštjan BAJEC/Valentin BUCIK/Klas Matija BRENK/Jagoda VATOVEC/Miha ŽARGI (2007) *Razvoj preizkusov procesiranja govornih dražljajev:kognitivnopsihološki in avdiološkividiki* [Development of speech audiometry tests: Cognitive psychological and audiological perspective]. Research report L5-6240. Ljubljana: Slovenian Research Agency.

PODLESEK, Anja/Luka KOMIDAR/Gregor SOČAN/Boštjan BAJEC/Valentin BUCIK/Klas Matija BRENK/Jagoda VATOVEC/Miha ŽARGI (2008) "A comparative analysis of different procedures for measuring speech recognition threshold in quiet." *Psihološka obzorja/Horizons of Psychology* 17/4, 33–49.

POMPE, Janko (1968) *Razvoj avdiometrije na ORL kliniki v Ljubljani* [Development of audiometry at ORL Clinic in Ljubljana]. Unpublished manuscript. Ljubljana: University Medical Center.

VAN DIJK, Johannes E/Jeff DUIJNDAM/Kees GRAAMANS (2000) "Acoustic neuroma: deterioration of speech discrimination related to thresholds in pure-tone audiometry." *Acta Oto-Laryngologica* 120/5, 627–632.

VERDONIK, Darinka/Iztok KOSEM/Ana ZWITTER VITEZ/Simon KREK/Marko STABEJ (2013) "Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS." *Language resources and evaluation* 47(4), 1031–1048.

ZEMLJARIČ MIKLAVČIČ, Jana/Marko STABEJ/Simon KREK/Ana ZWITTER VITEZ (2009) "Kaj in zakaj v referenčni govorni korpus slovenščine" In: M. Stabej (ed.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 437–442.

Reference Books and Corpora

GIGAFIDA – WRITTEN CORPUS, corpus of written Slovene: http://www.gigafida.net/

GOS – SPOKEN CORPUS, corpus of spoken Slovene: http://www.korpus-gos.net/

SLOVAR SLOVENSKEGA KNJIŽNEGA JEZIKA [Dictionary of Standard Slovenian] (2014), A. Bajec (ed.). Ljubljana: SAZU and Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika.

SLOVENSKI PRAVOPIS [Slovenian Ortography] (2014), J. Toporišič (ed.). Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. http://www.fran.si/134/slovenski-pravopis

Abstract
ADAPTING THE FREIBURG MONOSYLLABIC WORD TEST FOR SLOVENIAN

Speech audiometry is one of the standard methods used to diagnose the type of hearing loss and to assess the communication function of the patient by determining the level of the patient's ability to understand and repeat words presented to him or her in a hearing test. For this purpose, the Slovenian adaptation of the German tests developed by Hahlbrock (1953, 1960) – the Freiburg Monosyllabic Word Test and the Freiburg Number Test – are used in Slovenia (adapted in 1968 by Pompe). In this paper we focus on the Freiburg Monosyllabic Word Test for Slovenian, which has been criticized by patients as well as in the literature for the unequal difficulty and frequency of the words, with many of these being extremely rare or even obsolete. Since part of the patient's communication function is retrieving the meaning of individual words by guessing, the less frequent and consequently less familiar words do not contribute to reliable testing results. We therefore adapt the test by identifying and removing such words and supplement them with phonetically similar words to preserve the phonetic balance of the list. The words used for replacement are extracted from the written corpus of Slovenian Gigafida and the spoken corpus of Slovenian GOS, while the optimal combinations of words are established by using computational algorithms.

Keywords: speech audiometry, Freiburg Word Test, test adaptation, corpora

Povzetek
PRIREDBA FREIBURŠKEGA ENOZLOŽNEGA BESEDNEGA PREIZKUSA ZA SLOVENŠČINO

Govorna avdiometrija je eden od standardnih diagnostičnih pripomočkov pri ugotavljanju različnih tipov slušnega primanjkljaja ter pri preverjanju sporazumevalne funkcije pri pacientu, kjer s pomočjo testov slušne zaznave preverjamo, kakšna je pacientova zmožnost razumeti in ponoviti besede iz testa. V Sloveniji je v rabi Freiburški govorni preizkus (enozložni besedni in številčni preizkus), ki ga je razvil Hahlbrock (Hahlbrock 1953, 1960), za slovenske govorce pa leta 1968 priredil Pompe. V članku se osredotočimo na enozložni besedni preizkus, za katerega je bilo ugotovljeno veliko pomanjkljivosti predvsem z vidika pogostosti besed, saj test vsebuje kar precejšnje število izjemno redkih ali celo zastarelih besed. Ker je del sporazumevalne funkcije pri govorcu tudi zmožnost ugibanja slišane besede, je pri velikem številu govorcu neznanih besed pod vprašajem veljavnost izmerjenega rezultata, saj neznane besede govorec težje ugane. Test prenovimo tako, da najprej identificiramo manj pogoste in zastarele besede ter jih zamenjamo s fonetično podobnimi besedami, da obdržimo fonetično uravnoteženost testa. Nadomestne besede poiščemo z uporabo pisnega korpusa slovenščine Gigafida ter korpusa govorjene slovenščine GOS. Najbolj ustrezno kombinacijo nadomestnih besed, ki ohranja fonetično uravnoteženost testa, določimo z uporabo računskih algoritmov.

Ključne besede: govorna avdiometrija, freiburški govorni preizkus, priredba preizkusa, korpusi